IOP Conf. Series: Journal of Physics: Conf. Series **913** (2017) 012005 doi:10.1088/1742-6596/913/1/012005

Use of machine learning methods to classify Universities based on the income structure

Alexandra Terlyga[1] and Igor Balk[2]

[1] Ural Federal University after the first President of Russia B.N. Eltsin, 61 Komsomolskaya st., Ekaterinburg, Russian Federation.
[2] Global Innovation Labs LLC, 258 harvard Str #352 Brookline MA 02446 USA

E-mail: info@innovationlabs.net

Abstract. In this paper we discuss use of machine learning methods such as self organizing maps, k-means and Ward's clustering to perform classification of universities based on their income. This classification will allow us to quantitate classification of universities as teaching, research, entrepreneur, etc. which is important tool for government, corporations and general public alike in setting expectation and selecting universities to achieve different goals.

1. Introduction

In the modern society Universities become the key factor of economic development, because they function not only as establishments training staff and leading scientific research but also become active participants of innovation processes in Economics [1]. University should make its contribution in development of regional and national economics by creating the system of transfer of the research results into industry, Etzkowitz named it "Triple Helix system of innovation".

From the end of 1990s much attention to the concept of entrepreneurial university is given by Academic society, which tries to define this phenomenon in its works [2–9]. In the abovementioned works the definition of entrepreneurial university is given with the help of qualitative characteristics set determining the university activity in the sphere of business studies, start-up formation and organization of the projects with enterprises on technology transfer to industry providing.

Thus problems of universities financial activity become important subject for consideration as they help to assess incomes structure with the purpose to find out the most valuable sources and taking management decisions in the direction of university to the entrepreneurial type. What criteria should be concerned while university entrepreneurial activity assessing? How during analyzing the structure of University incomes, development can be managed?

In the number of modern works there is an attempt to analyze university financial indicators and to promote offers on changes of the current state and structural alterations. Concepts and mechanisms of financial management by universities need innovations in connection with changes in the system of market economy and corresponding financial crisis [10]. Analysis of the situation with expenses at the University supposes forming the solution on budget management strengthening, in particular on financial assessment indicators system creation [11] Issues on approach to efficiency assessment of governmental means into research university infrastructure, regional innovation systems and public infrastructures are discussed [12]. Current condition of university finance management in Czech Republic is analyzed and offers on universities budget assignations allocation are performed in connection with required goals at the State University [13]. Comparing structures of incomes

and expenditures of Canadian and Ukrainian Universities offers on investments from regional and municipal budgets are represented [14]. The role of accounting information and its influence on organizational changes from the viewpoint of competences and staff responsibilities becomes more valuable [15]. Analysis of budgeting system in four different universities from four ethnical groups allows to make recommendations on accounting cultural diversity in education [16].

In this paper we will concentrate on study of university classification using machine learning clustering algorithms.

2. Methodology

In this study we used list of universities according to their ranking by U.S. News Best Global Universities. This ranking is mainly focused on the academic research done in the University and worldwide reputation. For the ranking it utilize the daft by Clarivate Analytics (earlier Thomson Reuters), and accordingly Web of Science indexing service [17, 18].

To obtain information about income sources we used data provided in consolidated reports of the US Universities. These reports have uniform structure appropriate for performing further structuring, compilation and clustering.

The original US News ranking consisted of 170 Universities. The first hundred were chosen for this study. To unify reports analysis of universities income items was performed and the structure, which most relevantly describes major articles of universities income was created.

Incomes from educational activity are divided into revenues from students training in major education programs and earnings from additional educational programs:

1) Net student Income
2) Continuing education and executive programs
3) Total sponsored support is divided into the following categories:
4) Federal grants
5) State, Local grants
6) Private grants
7) Recovery of indirect costs and allowances

Total research revenues is divided into

1) Research revenues
2) Special laboratories revenue

Revenues from enterprises founded to provide supplementary university infrastructure activity are presented in the item:

- Auxiliary enterprises

Gifts for current use and other types of charity are shown in the item:

- Contributions and gifts

The following items present investment activities of the Universities:

1) Endowment returns made available for operations
2) General Operation Activities(GOA) returns made available for operations
3) Other investment income

Other incomes which do not have uniform structure and depend on peculiarities of the university activity and which are shown in the report with the aggregative line, presented in the item:

- Other incomes.

According to the chosen structure of university revenues the unified database was established. Data was additionally normalized:

1) to the total number of students educated at the University.
2) to the total income of the University.

3. Overview of the clustering algorithms

In this study we have used Ward's Hierarchical Agglomerative Clustering Method, K-means Method and Self-organizing map (SOM).

First we have established existence of clusters using SOM. The Self-Organizing Map (SOM), commonly also known as Kohonen network [19, 20] is a computational method for the visualization and analysis of high-dimensional data, especially experimentally acquired information.

The Self-Organizing Map defines an ordered mapping, a kind of projection from a set of given data items onto a regular, usually two-dimensional grid. A model mi is associated with each grid node. These models are computed by the SOM algorithm. A data item will be mapped into the node whose model which is most similar to the data item, e.g., has the smallest distance from the data item in some metric.

Algorithm

1. Randomize the map's nodes' weight vectors
2. Grab an input vector D(t), which is a target input data vector.
3. Traverse each node in the map
    1. Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector
    2. Track the node that produces the smallest distance (this node is the best matching unit, BMU)
4. Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector

$$W_v(s+1) = W_v(s) + \theta(u, v, s) * \alpha(s) * (D(t) - W_v(s))$$

where

s – is the current iteration

λ – is the iteration limit

t – is the index of the target input data vector in the input data set D

D(t) – is a target input data vector

v – is the index of the node in the map

W_v – is the current weight vector of node v

u – is the index of the best matching unit (BMU) in the map

θ(u, v, s) – is a restraint due to distance from BMU, usually called the neighborhood function, and

α (s) – is a learning restraint due to iteration progress.

5. Increase s and repeat from step 2 while s< λ

Later we used Ward's clustering and k-means algorithm to confirm stability of the obtained clustering model and study it's structure.

Ward's Hierarchical Agglomerative Clustering Method was first introduced by Ward [21]. He suggested an agglomerative hierarchical clustering algorithm with the criterion for choosing the pair of clusters to merge at each level is defined by the optimal value of an objective function.

We define a distance as a positive, definite, symmetric mapping of a pair of observation vectors onto the positive real, which in addition satisfies the triangular inequality.

For observations i, j, k we have:

$$d(i, j) > 0; d(i, j) = 0 \Longleftrightarrow i = j;\ d(i, j) = d(j, i); d(i, j) \leq d(i, k) + d(k, j).$$

For an observation set, I, with i,j,k ∈ I we can write the distance as a mapping from the Cartesian product of the observation set into the positive real : $d : I \times I \rightarrow R+$.

At each step of the clustering algorithm we find a pair of clusters that leads to a minimum increase in total within cluster variance after the merging i.e. weighted squared distance between cluster centers $(1/|q| \sum_{i \in q} d^2(i, q^*))$.

Initially all clusters contain only a singe point. We can define Euclidean distance squared using norm ‖.‖: if i,i′ ∈ R|J|, i.e. these observations have values on attributes j ∈ {1, 2, . . . , |J |}, J is the attribute set, |.| denotes cardinality, then

$$d^2(i, i') = \|i - i'\|2 = \sum_j (i_j - i'_j)^2$$

Suppose that clusters C_i and $C_J$were next to be merged. At this point all of the current pairwise cluster distances are known. The recursive formula gives the updated cluster distances following the pending merge of clusters C_i and C_J. Let

•$d_{ij}, d_{ik}\ and\ d_{jk}$ be the pairwise distances between clusters $C_i, C_j\ and C_k$, respectively,

• $d_{(ij)k}$be the distance between the new cluster $C_i \cup\ C_j and\ C_k$.

An algorithm belongs to the Lance-Williams family if the updated cluster distance $d_{(ij)k}$ can be computed recursively by

$$\mathrm{d_{(ij)k}} = \alpha_\mathrm{i}\mathrm{d_{ik}} + \alpha_\mathrm{j}\mathrm{d_{jk}} + \beta\mathrm{d_{ij}} + \gamma\left|\mathrm{d_{ik}} - \mathrm{d_{jk}}\right|,$$

where $\alpha_i, \alpha_j, \beta\ and\ \gamma$ are parameters, which may depend on cluster sizes, that together with the cluster distance function d_{ij}determine the clustering algorithm.

The term "k-means" was first used by James MacQueen in 1967 [22] though the idea goes back to Hugo Steinhaus in 1957 [23]. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published outside of Bell Labs until 1982 [24].

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = $\{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\arg\min_{\mathrm{S}} \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{x}\in\mathrm{S_i}} \|\mathrm{x} - \mu_\mathrm{i}\|^2 = \arg\min_{\mathrm{S}} \sum_{\mathrm{i}=1}^{\mathrm{k}} |\mathrm{S_i}|\mathrm{Var}\mathrm{S_i}$$

where μ_i is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\sum_{\mathrm{Cluster\ C_i}} \sum_{\mathrm{Dimension\ d}} \sum_{\mathrm{x,y}\in\mathrm{C_i}} (\mathrm{x_d} - \mathrm{y_d})^2$$

Because the total variance is constant, this is also equivalent to maximizing the squared deviations between points in different clusters (between-cluster sum of squares, BCSS) [25].

Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$, the algorithm proceeds by alternating between two steps [26].

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.

$$\mathrm{S_i^{(t)}} = \left\{\mathrm{x_p}: \left\|\mathrm{x_p} - \mathrm{m_i^{(t)}}\right\|^2 \le \left\|\mathrm{x_p} - \mathrm{m_j^{(t)}}\right\|^2\ \forall\ \mathrm{j};\ 1 \le \mathrm{j} \le \mathrm{k}\right\}$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$\mathrm{m_i^{(t+1)}} = \frac{1}{\left|\mathrm{S_i^{(t)}}\right|} \sum_{\mathrm{x_j}\in\mathrm{S_i^{(t)}}} \mathrm{x_j}$$

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitioning's, the algorithm must

converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm.

The algorithm is often presented as assigning objects to the nearest cluster by distance. The standard algorithm aims at minimizing the WCSS objective, and thus assigns by "least sum of squares", which is exactly equivalent to assigning by the smallest Euclidean distance.

4. Empirical results

As a first step of our investigation we created 3 x 3 SOM so see the structure of the data and that cluster analysis will make sense in this case. As can be seen from the Figure 1 this approach worked very well and we have clear separation of two large groups of the universities (cells with centers around (0, 175), (1, 175) on one side and (1, 0), (2, 0) and (2.5, 1)). Now we can use k-means and Ward's clustering algorithm to validate our approach.

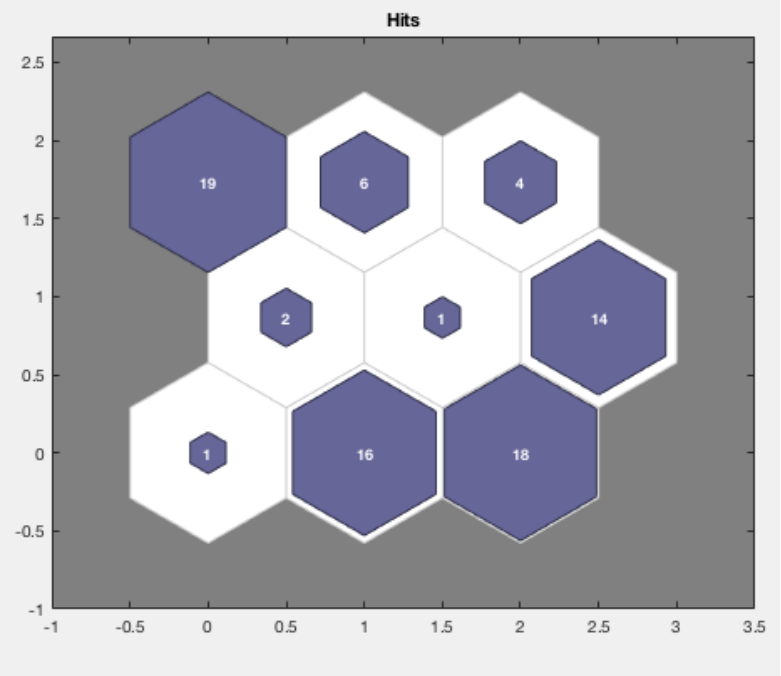

Figure 1. SOM hits with 3 x 3 mapping

As a check of validity of our clustering model we used silhouette metric. The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the i-th point, Si, is defined as Si = (bi-ai)/ max(ai,bi) where ai is the average distance from the i-th point to the other points in the same cluster as i, and bi is the minimum average distance from the i-th point to points in a different cluster, minimized over clusters. The silhouette value ranges from -1 to +1. A high silhouette value indicates that i is well-matched to its own cluster, and poorly-matched to neighboring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution may have either too many or too few clusters. The silhouette clustering evaluation criterion can be used with any distance metric. On the figures 2 and 3 we can see silhouette plot of k-means and Ward's clustering. We used simple squared Euclidean distances to calculate silhouette. As we can clear see from this figures most of the data gives good fit within its clusters especially with k-means clustering, with Ward's clustering we see that third cluster is really artificial and we should stick with two clusters approach using this algorithm.

Our next goal was to check how accurate is our clustering i.e. are we getting same clusters from the both algorithms. In order to do so we used Matlab crosstab function, which returned chi-squared

value of 157.4737 for Pearson's chi-squared test of independence and corresponding p-value $5.0896*10^{-33}$, which means strong correlations between clusters obtained by both algorithms.

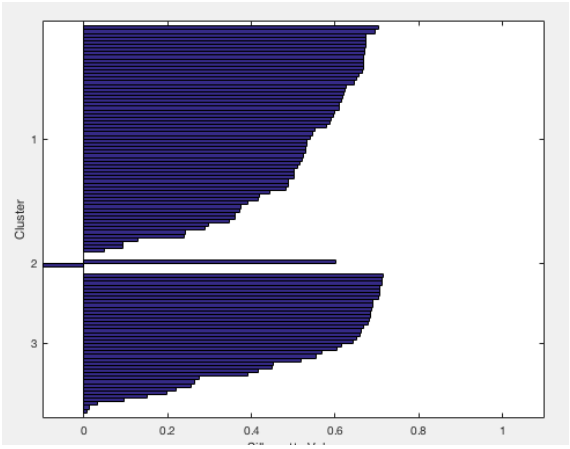

Figure 2. K-means clustering silhouette values

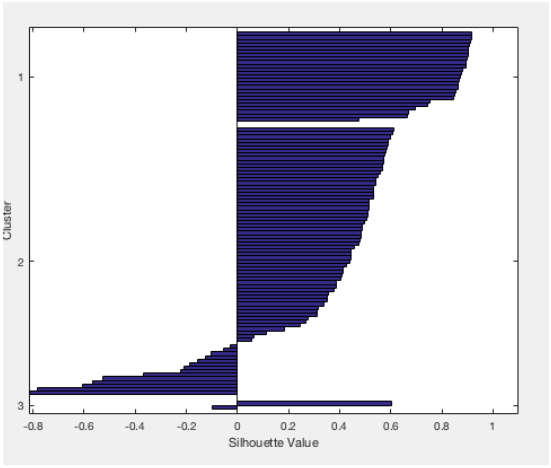

Figure 3. Ward clustering silhouette values

5. Conclusion

In this study we have clearly show that machine learning clustering techniques can be used to perform classification of the universities by the source of their income. The next logical step would be to cluster universities by their spending and overall P&L statements. This classification will allow us to quantitate classification of universities as teaching, research, entrepreneur, etc. which is important tool for government, corporations and general public alike in setting expectation and selecting universities to achieve different goals.

References

[1] Etzkowitz H 2003 Research groups as 'quasi-firms': the invention of the enterpreneurial universiry *Research Policy* **32** p 109–121
[2] Clark B 1998 *Creating enterpreneurial universities: organizational pathways of transformation* (New York)
[3] Etzkowitz H 1998 The norms of enterpreneurial science: cognitive effects of the new university – industry linkages *Research Policy* **27(8)** p 823–833
[4] Gibb A 2005 Towards the Enterpreneurial University: Enterpreneurship Education as a level for change [online] Avaliable from: http://ncee.org.uk/wp-content/uploads/2014/06/towards_the_entrepreneurial_university.pdf
[5] Guerrero M and Urbano D 2012 The development of an entrepreneurial unovercity *The Journal of Technology Transfed* **37(1)** p 43–74
[6] Gulbransden M and Slipersater S 2007 The third mission and the enterpreneurial university model, ed A. Bonaccorsi *Universities and Strategic Knowledge Creation*: *Specialization and performance In Europe* (Edward Elgar Publishing) p 112–144
[7] Kirdy D A 2006 Creating enterpreneurial universities in the UK: applying enterpreneurship theory to practice *Journal of Technology Transfer* **31(5)** p 599–603
[8] Montesinos P et al 2008 Thid Mission Ranking for World Class Universities: Beyond Teaching and Research *Higher Education in Eroupe* **33(2–3)** p 259–271
[9] Nelles J and Vorley T 2010 Constructing an Enterpreneurial Architecture: An Emergent Framework for Studing the Contemporary University beyond the Enterpreneurial Turn *Innovative Higher Education* **35(3)** p 161–176

[10] RETRACTED ARTICLE: University financial management concepts and mechanisms of innovation Sun (M.-C. 2010 Proceedings 2010) *2nd IEEE International Conference on Information and Financial Engineering* ICIFE 2010 5609412 p 517–521
[11] Wu C and Wang W 2011 University financial revenues and expenditures of the existing problems and countermeasures *2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC 2011)* p 6895–6898
[12] Scaringella L and Chanaron J-J 2016 Grenoble – GIANT Territorial Innovation Models: Are investments in research infrastructures worthwhile? *Technological Forecasting and Social Change* **112** p 92–101
[13] Zonca P, Moravcova V, Maresova P and Kuca K 2015 Financing of a Public Universities in Czech Republic *Proceedings of the 25th International Business Information Management Association Conference – Innovation Vision 2020: From Regional Development Sustainability to Global Economic Growth (IBIMA 2015)* p 198–201
[14] Totska O L 2014 Structure of university revenues and expenditures in Canada and Ukraine: Comparative analysis *Actual Problems of Economics* **151(1)** p 225–232
[15] Bonollo E, Lazzini S and Zuccardi Merli M 2017 Accounting information system and organizational change: An analysis in "first mover" public Universities *Lecture Notes in Information Systems and Organisation* p 183–201
[16] Ponorîcă A G, Popa A F and Stănilă G O 2013 Comparative study of budgeting systems within various European universities *Quality – Access to Success 14* (SUPPL.2), p 53–65
[17] *Best Global Universities* Avaliable from: https://www.usnews.com/education/best-global-universities
[18] *U.S. Colleges and Universities-Utility Patent Grants, Calendar Years 1969–2012* Avaliable from: https://www.uspto.gov/web/offices/ac/ido/oeip/taf/univ/univ_toc.htm
[19] Kohonen T 1982 Self-organized formation of topologically correct feature aps *Biological Cybernetics* **43** p 59–69
[20] Kohonen T 2001 Self-Organizing Maps *Third, extended edition* (Springer)
[21] Ward J H, Jr 1963 Hierarchical Grouping to Optimize an Objective Function *Journal of the American Statistical Association* **58(301)** p 236–244
[22] MacQueen J B 1967 Some Methods for classification and Analysis of Multivariate Observations *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press) p 281–297
[23] Steinhaus H 1957 Sur la division des corps matériels en parties *Bull. Acad. Polon. Sci.* (in French) **4(12)** p 801–804
[24] Lloyd S P 1957 Least squares quantization in PCM *IEEE Transactions on Information Theory* **28(2)** p 129–137
[25] Kriegel H-P, Schubert E and Zimek A 2016 The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems* **52(2)** p 341–378
[26] MacKay D 2003 *Information Theory, Inference and Learning Algorithms* (UK, Cambridge University Press)